



Arbres et généalogie politique. Représentation arborée du discours de Jacques Chirac (1995-2002)

Damon Mayaffre, Xuan Luong

► To cite this version:

Damon Mayaffre, Xuan Luong. Arbres et généalogie politique. Représentation arborée du discours de Jacques Chirac (1995-2002). *Histoire & Mesure*, 2003, XVIII (3/4), pp.289-311. hal-00551595

HAL Id: hal-00551595

<https://hal.science/hal-00551595>

Submitted on 4 Jan 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

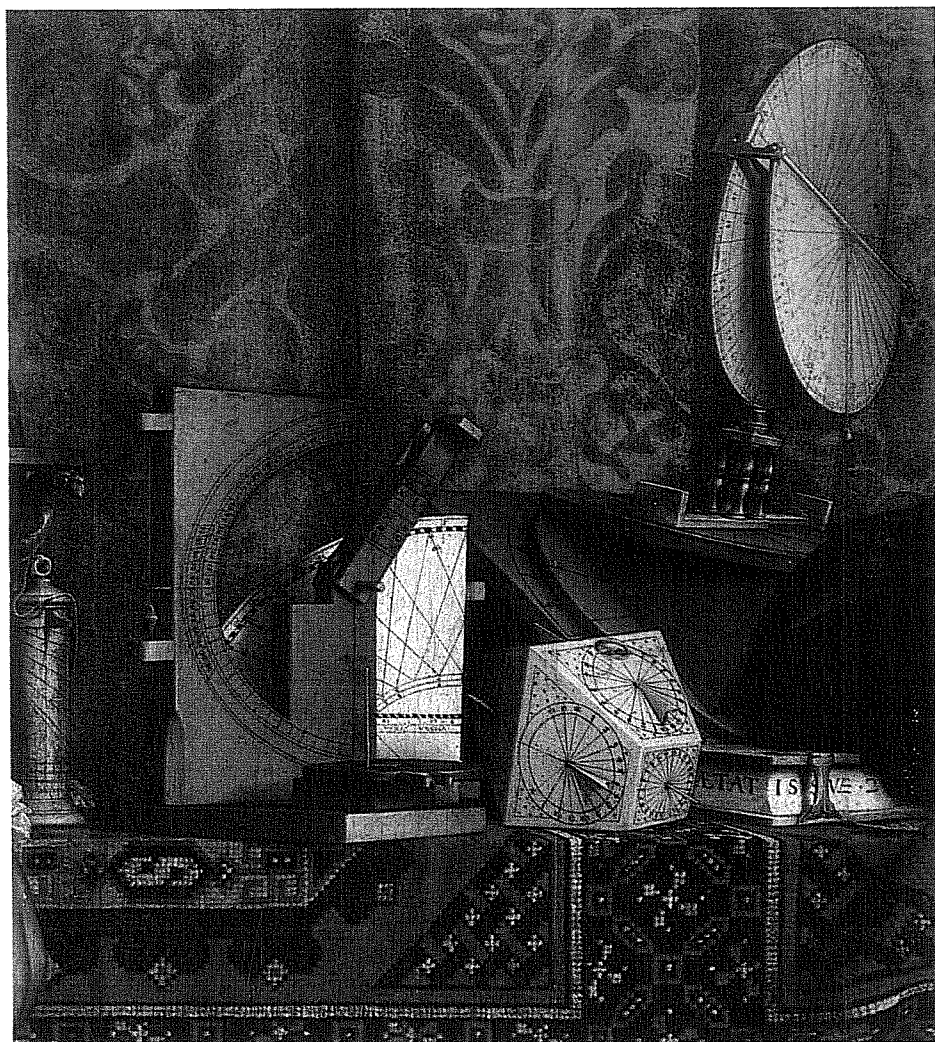
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HISTOIRE & MESURE

2003

VOLUME XVIII — n° 3/4

MESURER LE TEXTE



Damon Mayaffre & Xuan Luong*

**Les discours de Jacques Chirac (1995-2002).
Représentation arborée et généalogie politique**

Fac-similé

Résumé. L'analyse arborée fait appel d'une part aux techniques classificatoires et, d'autre part, à la topologie mathématique. Elle permet d'organiser, de classer, de situer les textes sur un arbre pour représenter au mieux la distance de chacun d'entre eux à tous les autres.

Sur un important corpus de textes présidentiels (1958-2002), elle donne à voir l'absence de généalogie politique de J. Chirac. Sur un corpus de textes plus modeste (1997-2002), elle donne à réfléchir sur la victoire du président sortant et sur la défaite de L. Jospin.

Abstract. Tree-Analysis and political genealogy of J. Chirac. Tree-analysis enlists both classification techniques and mathematical topology. It makes it possible to organise, classify and situate the texts of a given corpus in order to provide the best representation of the distance between each one and all the others.

Applied to an important body of texts by French presidents it highlights J. Chirac's lack of political ancestry. Applied to a more modest corpus (1997-2002) it gives food for thought about the previous president's victory and L. Jospin's defeat.

* CNRS, UMR 6039, Bases, Corpus et Langage, UFR Lettres, Arts et Sciences humaines, 98 bd Édouard Herriot, BP 3209 —Nice Cedex. E-mail : mayaffre@unice.fr et luong@unice.fr

Identifier et classer ses objets : une exigence pour toutes les études en sciences naturelles et humaines ; une nécessité par exemple pour l'historien qui analyse un grand nombre de discours politiques.

Soit deux textes politiques : le but est de les comparer, terme à terme, pour définir leur originalité respective. Soit cinq, dix, cent textes : l'enjeu devient de les classer, c'est-à-dire de les grouper en classes d'objets voisins pour avoir une vision ordonnée et synthétique du corpus et tenter de cerner, de manière dynamique, les spécificités individuelles de ses composantes.

L'analyse du discours présidentiel sous la v^e République de ses origines à nos jours (1958-2002) n'échappe pas à cette préoccupation classificatoire, que l'on appellera ici « taxinomique », ailleurs « typologique ». Nous disposons, à l'intérieur du corpus, de six grands Textes (en réalité six sous-corpus de textes apparentés) : les discours produits par les présidents durant les six septennats que compte la v^e République. Et les principales interrogations historiques face à ce corpus seront les suivantes : comment se situe, par exemple, le discours de Jacques Chirac (1995-2002) par rapport à celui de ses prédécesseurs ? Le discours chiraquien est-il un discours gaulliste proche de celui de de Gaulle ou de Pompidou, ou un discours orléaniste proche de celui de Giscard d'Estaing ? Ou encore : les deux sous-corpus Mitterrand (1981-1988 et 1988-1995) restent-ils jumelés à l'intérieur du corpus présidentiel, malgré la trajectoire politique du président de gauche durant ses deux septennats, et contrastent-ils vraiment, ensemble, avec les sous-corpus des présidents de droite ?

L'analyse du discours de la cohabitation Jospin/Chirac (1997-2002) n'échappe pas non plus à cette exigence de classification. Nous analyserons douze sous-corpus composés des discours annuels de Jospin et de Chirac entre 1997 et 2002 (Jospin-1997, Jospin-1998, ..., Jospin-2002 et Chirac-1997, Chirac-1998, ..., Chirac-2002). Et les interrogations seront du type : existe-t-il un dénominateur commun important entre les différents discours annuels de Jospin versus les différents discours annuels de Chirac ? En d'autres termes, le Président et le Premier ministre ont-ils une personnalité politique et discursive qui résiste à la chronologie et qui fait que, quel que soit le thème abordé ou la conjoncture politique du moment, un discours de Jospin reste apparenté aux autres discours de Jospin pour s'éloigner des discours de Chirac ? Par ailleurs, comment, à l'intérieur de leur discours respectif, la personnalité politico-linguistique de Jospin et de Chirac évolue-t-elle entre 1997 et 2002 ? Évolution progressive ? Dans ce

cas les discours de 1997, par exemple, seront à proximité de ceux de 1998 ou 1999, et beaucoup plus éloignés de ceux de 2002. Évolution chaotique ? Dans ce cas l'analyse devra montrer les rapprochements et les ruptures inattendus.

La logométrie se trouve donc ici investie de deux tâches d'importance, successives et complémentaires : d'une part *calculer* la distance entre différents textes, d'autre part, *représenter* par un graphique cette distance pour la rendre visible, palpable, interprétable pour l'analyste.

Après un bref rappel des enjeux statistiques et linguistiques du calcul de la distance intertextuelle, nous le supposerons connu, pour exposer un outil de représentation encore peu exploité en Histoire : l'*analyse arborée*¹, aujourd'hui partie intégrante du logiciel de logométrie *Hyperbase* conçu par Étienne Brunet et produit par l'UMR 6039, Bases, Corpus et Langage (Nice).

1. Arbre et généalogie de J. Chirac

Le premier mouvement de cette contribution a une finalité méthodologique mais il s'appuie sur un corpus réel et un exemple concret qui intéresseront l'historien : l'analyse arborée doit en effet nous éclairer sur le discours présidentiel français en général (1958-2002) et sur la parole chiraquienne en particulier. Le but ici est de ne pas opposer méthode et résultats, analyse formelle et interprétation, logométrie et Histoire.

Le corpus présidentiel et sa représentation arborée

Le corpus de discours étudiés compte les principales interventions télévisées des présidents de la République depuis 1958. En termes quantitatifs, il se présente tel qu'il est décrit dans le Tableau 1.

Le corpus est homogène quant au genre de discours et équilibré, entre les locuteurs, quant au nombre de discours. Surtout le corpus est pluriel ou contrastif. Certes, il n'est composé que de six parties (ou six Textes), aussi serait-il sans doute envisageable de faire l'examen successif des particula-

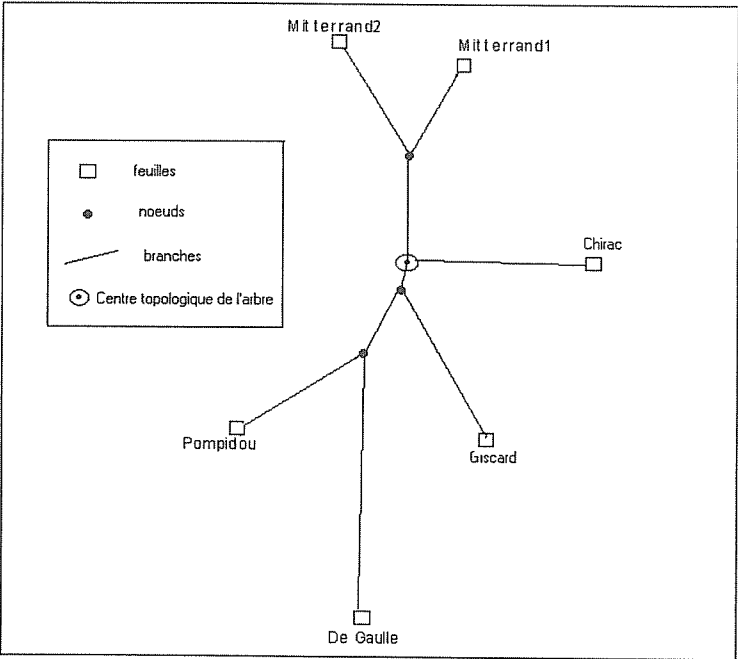
1. L'analyse arborée que nous présentons ici a été mise au point par Xuan Luong et Jean-Pierre Barthélemy. Elle a été décrite en 1988 dans la thèse d'État de X. LUONG, 1988 à laquelle nous renvoyons pour la théorie mathématique. Elle a été expliquée aux mathématiciens et aux littéraires dans divers articles de vulgarisation cités dans cet article et recensés en bibliographie.

rités de chacune d’entre elles pour avoir une vision détaillée de l’ensemble. Mais ce sont des questions de synthèse que nous voulons d’emblée poser : quels degrés de ressemblance ou au contraire de dissimilarité existe-t-il entre ces différents Textes ? Finalement, y-a-t-il un *ordre* dans le corpus (ordre chronologique ? ordre politique ?) qui se dégage et qui rapprocherait tel président de tel autre ou éloignerait celui-ci de celui-là ?

Tableau 1. *Le corpus présidentiel (1958-2002)*

Textes	Nombre de discours	Nombre d'occurrences ²	Nombre de vocables différents
Texte 1 : de Gaulle (1958-1969)	79	222 090	11 949
Texte 2 : Pompidou (1969-1974)	112	223 182	13 076
Texte 3 : Giscard (1974-1981)	116	406 618	13 898
Texte 4 : Mitterrand 1 (1981-1988)	70	367 067	14 849
Texte 5 : Mitterrand 2 (1988-1995)	85	336 980	15 273
Texte 6 : Chirac (1995-2002)	111	337 536	13 242

Figure 1. *Analyse arborée du corpus présidentiel (1958-2002) selon le vocabulaire*



2 « Occurrences » est pris ici dans le sens élargi des informaticiens, ils s’agit non seulement de tous les mots graphiques mais aussi de toutes les ponctuations.

L'ambition est donc de mesurer les distances ou les connexions entre les six Textes des six septennats, et de (se) représenter ces distances pour pouvoir les interpréter. L'on se propose de voir immédiatement le résultat graphique que produit l'analyse arborée du vocabulaire des discours (Figure 1).

Grossièrement, l'arbre se lit et s'interprète comme suit. Il y a sous la v^e République deux grands types de discours qui du point de vue du vocabulaire employé s'opposent : le discours que l'on qualifiera de gaulliste situé sur une branche à une extrémité de l'arbre, et le discours que l'on qualifiera de miterrandien situé sur la branche opposée à l'autre extrémité. Au bout de ces deux branches opposées, se dessine un nœud ou une ramification terminale à deux feuilles qui montrent, d'un côté, sur la branche gaulliste, que les discours de de Gaulle et de Pompidou sont apparentés et de l'autre, sur la branche miterrandienne, que les corpus Mitterrand 1 et Mitterrand 2 sont jumelés.

Le discours de Giscard d'Estaing et celui de Chirac apparaissent intermédiaires entre les deux branches. Dans le détail nous verrons qu'ils ne le sont pas au même titre (Giscard d'Estaing apparaissant affilié au discours gaulliste). Mais l'on peut déjà constater sans risque de se tromper que le discours de Jacques Chirac est, sur l'arbre, particulièrement central (il est rattaché au centre topologique de l'arbre), comme à mi-chemin des deux pôles, gaulliste et miterrandien. Le président, aux fins conseillers politiques, semble avoir réussi à faire une synthèse du vocabulaire de gauche et de droite, du vocabulaire classique et moderne, du gaullisme et du miterrandisme : son discours refuse ainsi de se marquer, de se démarquer, de s'opposer. Au centre (topologique de l'arbre), il apparaît susceptible de rallier tout le monde et de ne désobliger personne.

Cette lecture rapide de l'arbre et cette interprétation historique grossière demandent évidemment à être précisées. Elles n'ont d'autre ambition que d'introduire aux propos méthodologiques.

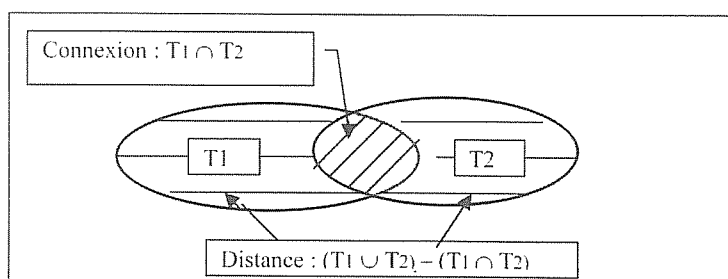
Le calcul de la distance

L'objet de cet article n'est pas de décrire le calcul de la distance intertextuelle dont l'analyse arborée, ici, donne une représentation. Depuis le commencement de la statistique lexicale une longue littérature y est consacrée³. Le dernier numéro de la revue *Corpus* y est entièrement dévoué⁴, et on y lira nécessairement l'article de Dominique Labbé : « La distance intertextuelle ». Il convient néanmoins de faire les rappels les plus

élémentaires pour mettre en perspective la présentation théorique de l'analyse arborée qui sera proposée.

À l'origine, dans les années 1970-1980, la « distance intertextuelle » d'aujourd'hui était communément envisagée sous l'angle de la « connexion lexicale ». Il s'agit là, en apparence, avant tout d'une évolution d'appellation et d'une simple différence de point de vue, selon que l'on insiste sur ce qui rapproche deux textes (connexion) ou complétement sur ce qui les éloigne (distance). Dans les deux cas, l'idée consiste à mesurer le vocabulaire commun à deux textes ou complétement le vocabulaire exclusif de l'un et de l'autre (Figure 2) ⁵.

Figure 2. *Connexion et distance entre textes*



Pourtant, derrière la mutation terminologique (« connexion lexicale » => « distance intertextuelle ») se cache une évolution plus profonde tant d'un point de vue linguistique que statistique.

Linguistiquement d'abord, le texte recouvre une réalité plus large que le lexique. Avec le développement des lemmatiseurs, il est aujourd'hui possible d'étudier non seulement les mots graphiques, mais aussi les lemmes, les catégories grammaticales ou encore les enchaînements syntaxiques. La distance intertextuelle actuelle peut donc être mesurée selon ces différents niveaux linguistiques lorsque la connexion lexicale ne prenait en compte, tout au moins au départ, que les formes graphiques.

Surtout, statistiquement, si les chercheurs sous l'impulsion de D. Labbé préfèrent parler de distance intertextuelle c'est que les bases du

3. MULLER, C., 1977, chap. 23, pp. 145-154 ; BRUNET, É., 1988 ; LABBÉ, D. & HUBBERT, P., 1998, pp. 361-370.

4. *Corpus*, n° 2, 2003. Notons aussi l'effort de synthèse de B. PINCEMIN, 2002, pp. 35-52.

5. Ce schéma nous a été directement inspiré par C. et D. LABBÉ, 2003.

calcul ont changé pour s'élargir à l'ensemble de la population du texte. Là où C. Muller ou É. Brunet ne prenaient en compte que la présence ou l'absence d'un terme dans deux textes pour mesurer leur connexion ⁶, on cherche dorénavant à considérer tous les termes selon leur fréquence ou leur poids dans les textes, c'est-à-dire leur répartition dans le corpus ; là où finalement on finissait par travailler avant tout sur les hapax (par définition présents dans un seul texte), on travaille aujourd'hui sur toute la surface du corpus, on prend en compte aujourd'hui tout le texte. Evidemment, l'algorithme qui en découle est plus complexe ⁷ notamment pour neutraliser les différences de taille entre les textes comparés ; car si la répartition des hapax dans les diverses parties du corpus répond à une simple règle de trois, la répartition des mots à fréquences élevées répond à des règles zipféennes plus compliquées ⁸.

Au final, le calcul complexe de la distance aboutit à un *indice* de distance très simple à lire compris entre 0 et 1. Égal à 0 (distance nulle), il signifie que les deux textes comparés sont en tous points identiques. Égal à 1 (distance maximale), cela signifie que les deux textes n'ont absolument rien en commun. Situé entre 0 et 1, cela signifie que les deux textes partagent un certain nombre de traits linguistiques (vocabulaire commun si la base du calcul est le vocabulaire, codes grammaticaux communs si l'on étudie la grammaire, etc.).

Sur un corpus comptant plusieurs textes, le calcul des distances permet d'établir un tableau matriciel carré dans lequel figure la distance, terme à terme, entre chacun des textes du corpus : la distance du texte 1 au texte 2, du texte 1 au texte 3, etc. Le tableau carré des distances entre textes du corpus présidentiel — la matrice de l'arbre que nous avons produit — figure dans le Tableau 2.

Un coup d'œil exercé pourrait ainsi remarquer que l'indice le plus faible (0,106) est celui qui sépare les deux corpus Mitterrand, preuve d'un vocabulaire en grande partie conjoint entre le premier septennat mitterrandien et le second. De la même manière, l'indice le plus élevé (0,229) est celui qui sépare de Gaulle de Mitterrand2. Au cas par cas, il serait ainsi possible de multiplier les constats sur ces comparaisons binaires ⁹.

6. Selon la méthode de P. JACQUART, 1908.

7. LABBÉ, C. et D., 2003.

8. Pour une discussion approfondie cf. C. et D. LABBÉ, 2003.

Tableau 2. *Indices des distances entre textes selon le vocabulaire dans le corpus présidentiel (1958-2002)*

	<i>de Gaulle</i>	<i>Pompidou</i>	<i>Giscard</i>	<i>Mitterrand1</i>	<i>Mitterrand2</i>	<i>Chirac</i>
de Gaulle	0	0,158	0,215	0,220	0,229	0,218
Pompidou	0,158	0	0,170	0,184	0,184	0,168
Giscard	0,215	0,170	0	0,184	0,178	0,159
Mitterrand1	0,220	0,184	0,184	0	0,106	0,164
Mitterrand2	0,229	0,184	0,178	0,106	0	0,151
Chirac	0,218	0,168	0,159	0,164	0,151	0

Mais le problème est d’avoir une vision synthétique d’un tel tableau (qui pourrait compter beaucoup plus de lignes et de colonnes selon le nombre de textes comparés). Car s’il est facile de lire et de se représenter la distance entre deux textes, il devient ardu à l’esprit humain de se représenter l’ensemble des distances qui séparent chacun des textes des autres et qui séparent tous de chacun.

L’analyse arborée

Précisément, l’analyse arborée se propose de résoudre ce problème en faisant appel d’une part aux techniques classificatoires classiques telles qu’on peut les voir fonctionner dans des dendogrammes traditionnels, et d’autre part à la topologie. Il s’agit d’organiser, de classer, de situer les textes sur un arbre pour représenter au mieux la distance de chacun d’entre eux par rapport à tous les autres.

Dans cette représentation, la proximité ou la réunion objective de deux textes (le vocabulaire partagé) est prise en compte, tout comme leur répulsion ou leur opposition communes à d’autres textes. C’est pourquoi il faut parler avec l’initiateur de la méthode d’une « dialectique regroupement/opposition »¹⁰ essentielle à la compréhension de l’analyse arborée ; le but étant de traiter ensemble et de réconcilier « les deux exigences, souvent contradictoires, de la classification automatique : homogénéité inter-classes (i.e : les regroupements) et séparation inter-classes (i.e. : les oppositions) »¹¹. En fait, dans un arbre, chaque paire de textes regroupés est

9. On aura évidemment remarqué que la distance d’un texte à lui-même est nulle, d’où la diagonale du tableau en 0.

10. BARTHÉLEMY, J.-P. & LUONG, X., 1998, p. 52.

11. BARTHÉLEMY, J.-P. & LUONG, X., 1998, p. 52.

opposée à toutes les autres paires de textes. Sur l'arbre présidentiel, la paire de Gaulle-Pompidou est opposée de manière évidente à la paire Mitterrand1-Mitterrand2, mais encore à la paire Mitterrand1-Chirac ou Mitterrand2-Giscard. Pour décrypter un arbre, il convient de repérer, sur le tracé plus ou moins sinueux du graphique, toutes ces structures minimales en « H » qui mettent en forme les regroupements de deux textes et leurs oppositions à d'autres couples de textes. De manière plus générale, lorsque l'on coupe une branche quelconque de l'arbre, le graphe se trouve partagé en deux parties ¹². Cette bipartition obtenue (et toutes celles possibles sur l'arbre) révèle les textes apparentés et tous ceux qui leur sont plus ou moins directement opposés (cf. Figure 3).

Structure et distance

C'est donc d'abord par sa structure (les regroupements représentés par des nœuds et les oppositions représentées par les branches) que se lit un arbre. C'est en ce sens que l'on a pu déchiffrer sur la Figure 1 l'existence de deux grands types de discours (ou deux branches maîtresses) opposés sous la v^e République (gaulliste/mitterrandien). Et, dans le même temps, que l'on a pu associer à l'intérieur de ces deux types de discours, les corpus de de Gaulle et de Pompidou *versus* les deux corpus Mitterrand : ils représentaient, respectivement, les deux feuilles d'une même branche en étant reliés par un nœud. On dit alors que les deux textes font partie de la même classe.

Évidemment, la structure de l'arbre dans notre exemple à six textes est riche de sens historique mais très élémentaire. On verra qu'elle peut être plus fine et plus complexe sur des corpus comptant un plus grand nombre de textes. La dialectique rapprochement/opposition s'exerce en effet non seulement sur la structure centrale de l'arbre (ses principales branches) mais localement sur toutes les parties ou sous-parties de l'arbre (ses différentes ramifications). Ainsi dans des corpus de plusieurs dizaines de textes, l'arborescence ou les ramifications se dessinent finement à plusieurs niveaux ¹³.

C'est ensuite par sa dimension ou par la longueur de ses branches que se lit un arbre. Plus précisément, la distance linguistique entre les textes

12. En termes mathématiques : deux parties connexes.

13. Cf., par exemple, de très beaux arbres dans M. JUILLARD, 1994, pp. 3-23 et X. LUONG, 1994, pp. 24-42. Les plus récents arbres ont été produits par M. KASTBERG SJÖBLOM, 2002, t. II, p. 461.

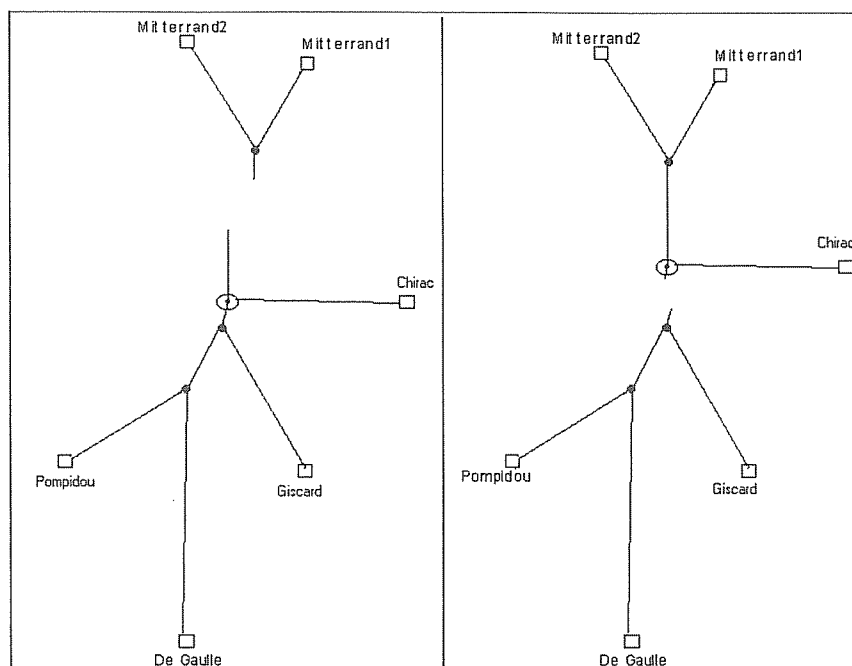
Figure 3. Exemples de bi-partitions de l'arbre présidentiel pour une aide à la lecture

Exemple 1

Cette partition de l'arbre en deux permet de montrer que la paire Mitterrand 1 et Mitterrand 2 est opposée à tous les autres textes

Exemple 2¹⁴

Cette partition de l'arbre en deux permet de montrer que le regroupement Giscard+(Pompidou-de Gaulle) est opposé aux textes de Chirac, Mitterrand1 et Mitterrand2



* La section de la branche Chirac montrerait que le texte de Chirac est apparenté à aucun autre texte et s'oppose à tous.

(qui représentent, en partie terminale des branches, les feuilles de l'arbre) est la distance géographique lue ou mesurée sur l'arbre en parcourant le chemin qui relie un texte à l'autre. Que l'on suive les branches qui relient le discours de de Gaulle à celui de Mitterrand2, par exemple, et l'on comprendra que leurs discours sont éloignés¹⁵ (donc dissemblables linguistiquement). Que l'on suive le chemin, plus court, qui mène de de Gaulle à Chirac et l'on devinera que leurs discours, sans mériter d'être regroupés

14. La section de la branche Chirac montrerait que le texte de Chirac n'est apparenté à aucun autre texte et « s'oppose » à tous.

15. Dans le détail mathématique, on notera que la distance entre de Gaulle et Mitterrand est une somme de distances ou distance additive des branches principales et des ramifications secondaires (J.-P. BARTHÉLEMY & X. LUONG, 1998, p. 50).

dans une même classe, sont plus ressemblants. Selon cette même logique comptable ou géométrique, lorsque l'on compare les deux regroupements opposés de Gaulle-Pompidou d'une part et Mitterrand1-Mitterrand2 d'autre part, l'on constate que le premier regroupement est plus lâche que le second : lue sur l'arbre, la distance de de Gaulle à Pompidou est plus importante que celle entre Mitterrand1 et Mitterrand2. Cela signifie que les deux textes de Mitterrand sont plus étroitement apparentés (ils partagent plus de vocabulaire commun et repoussent ensemble plus de mots) que ceux des deux premiers présidents de la V^e République.

À condition de ne jamais omettre de prendre en considération la structure de l'arbre, et notamment les nœuds ou regroupements qui marquent toujours les proximités prioritaires entre deux textes, cette représentation graphique de la distance apparaît donc très visuelle en matérialisant des indices mathématiques pas toujours parlants. Tout juste faut-il préciser que la distance construite sur le graphique par l'analyse arborée est obtenue par la meilleure « approximation » des distances réelles, calculées dans le tableau matriciel, mais en respectant les propriétés topologiques des graphes ¹⁶.

Dynamique de l'arbre

Un dernier aspect enfin de la construction des arbres, plus technique mais très éclairant, demande à être explicité.

La construction se fait de manière dynamique et progressive, mais avec une vision globale et synthétique de toutes les données du tableau matriciel. L'algorithme de X. Luong calcule d'abord les regroupements et les oppositions les plus marqués pour tracer les extrémités de l'arbre. Les premiers groupes sont ainsi formés ; ce sont les éléments topologiquement les plus structurés. Puis l'algorithme de construction de l'arbre ramène chaque groupe qui vient de se former à un représentant unique moyen pour examiner la nouvelle structure topologique et constituer, sur la base de ces nouvelles données, de nouveaux regroupements. Et ainsi de suite, jusqu'à épuisement, à force de réduction, des éléments du corpus ¹⁷.

16. BARTHÉLEMY, J.-P. & LUONG, X., 1998, p. 50.

17. Cette dynamique dans la construction des arbres explique pourquoi ils sont dits « non plantés » ou « non orientés » (contrairement aux arbres traditionnels produits par les classifications hiérarchiques, qui sont « plantés » par une racine). Ils se construisent au fur et à mesure en fonction des rapprochements et des oppositions les plus patents, d'abord, les plus fins ensuite, mais ne se structurent pas à partir d'une origine ou racine. De plus, si dans la

Par là, au fur et à mesure de la construction, se révèlent, hiérarchiquement, d'abord les attirances majeures entre les textes, puis de nouvelles affinités, secondaires mais très pertinentes, que l'on interprétera en terme d'intermédiarité ou de filiation. La figure 1, dans une nuance que nous n'avons pas encore relevée, souligne ainsi que le discours de Valéry Giscard d'Estaing est un discours intermédiaire entre le discours gaulliste et le discours miterrandien, mais il doit être identifié comme un discours du premier type ; il reste affilié au discours de de Gaulle et de Pompidou. Il se situe bien en effet sur la branche gaulliste même si, assez précocement, il s'en distingue par une ramification particulière très marquée : entre de Gaulle et Giscard, la sève, au départ, est commune (un vocabulaire de droite ?) même si leurs discours divergent. Soit dit encore autrement : sur la branche gaulliste, le discours giscardien apparaît, en termes de filiation, comme un lointain descendant du discours de de Gaulle¹⁸. Les caractères majeurs gaullistes — nettement identifiés dans le regroupement principal de Gaulle-Pompidou — sont sans doute en grande partie évanouis, mais quelques gènes encore ou quelques mots rappellent le discours originel¹⁹.

Précisons, enfin, d'une part que le logiciel fournit l'ordre d'apparition des regroupements pour souligner la hiérarchie de leur pertinence dans la structure de l'arbre : les regroupements de Gaulle-Pompidou et Mitterrand1-Mitterrand2 sont, on l'a compris dans notre exemple, les deux premiers regroupements (opposés) de l'arbre, le regroupement Giscard et la branche de-Gaulle-Pompidou vient seulement ensuite, il en est le troisième, etc.²⁰. D'autre part, un indice d'agrégation est calculé pour chaque nœud pour indiquer la fidélité de la représentation arborée par rapport au tableau matriciel : ici le regroupement de Gaulle-Pompidou

classification classique les éléments d'une même classe ont une même distance par rapport à chacun des autres éléments, ici, nous l'avons dit, la distance, sur l'arbre, entre deux textes varie selon la distance réelle calculée.

18. La lecture de l'arbre ne se fait pas d'une éventuelle origine ou racine vers l'extrémité des branches, mais des extrémités (qui constituent les regroupements les plus marqués) vers le centre topologique de l'arbre. C'est ainsi que le discours de Giscard d'Estaing est bien un rejeton du discours gaulliste.

19. En revanche, le discours de Chirac est purement intermédiaire entre les discours gaulliste et miterrandien. L'algorithme ne repère aucune filiation. Chirac se rattache directement, tel un rejeton bâtard, à la souche commune du discours politique, c'est-à-dire au centre topologique de l'arbre, et non à une branche maîtresse constituée, non à une famille politique référencée.

20. Pour saisir l'importance d'une vision dynamique de la construction des arbres, M. JUILLARD, 1994, pp. 18 et sq.

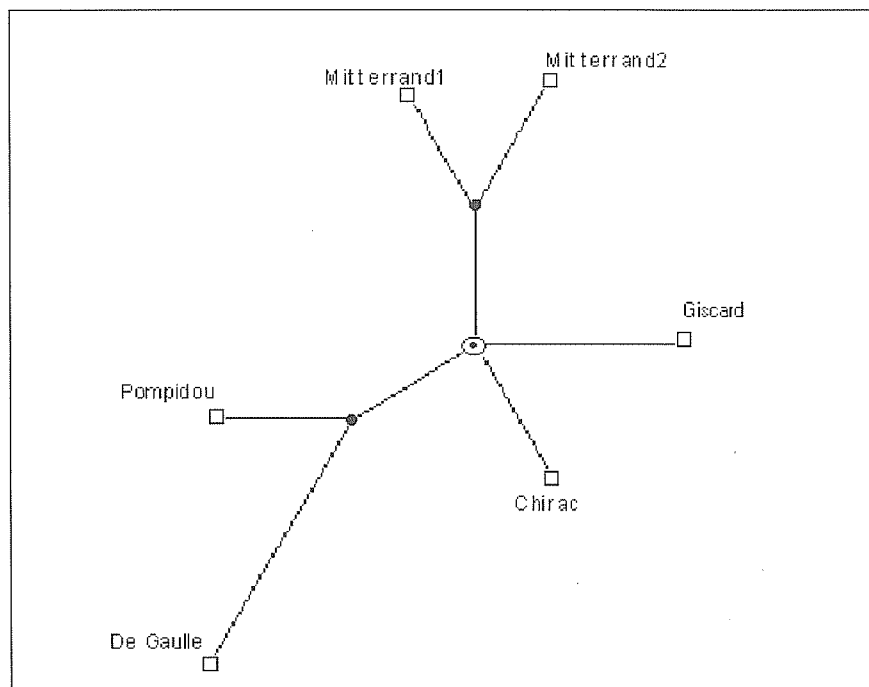
est affecté du coefficient maximal 1, c'est-à-dire qu'il est construit sans approximation des données brutes ; en revanche, le nœud reliant la ramification Giscard à la branche maîtresse qui mène à de Gaulle-Pompidou est lui affecté d'un coefficient d'agrégation moins important (0,95). Il signifie que 5 % des données ont été approximés pour constituer et représenter ce regroupement ²¹.

Retour au corpus présidentiel et à Jacques Chirac

Pour conclure ce premier mouvement méthodologique, nous voulons produire un deuxième graphique arboré représentant la distance entre les différents présidents non plus calculée sur le vocabulaire employé mais sur les codes grammaticaux utilisés (Figure 4).

La Figure 4 ressemble beaucoup à la Figure 1. Elle confirme ainsi l'analyse, nous permet de résumer la méthode et d'approfondir l'interprétation historique.

Figure 4. *Analyse arborée du corpus présidentiel (1958-2002) selon les catégories grammaticales*



21. LUONG, X., 1994, p. 33.

Comme sur la Figure 1, deux classes de textes s'opposent nettement : de Gaulle et Pompidou *versus* Mitterrand1 et Mitterrand2. Les codes grammaticaux responsables de ces regroupements/oppositions sont très intéressants même si nous ne pouvons pousser ici très loin l'analyse ²². Les discours de de Gaulle et de Pompidou apparaissent dominés par les noms, les adjectifs et les déterminants, lorsque les discours de Mitterrand sont dominés par les pronoms, les verbes et les adverbes. De la même manière qu'en littérature donc, deux discours typiques existent dans la rhétorique politique : le discours nominal et le discours verbal ²³.

Comme sur la Figure 1, le discours de Giscard d'Estaing est intermédiaire sans même rester affilié au discours nominal gaulliste. Cette situation d'intermédialité du discours giscardien trouve une explication simple : c'est une logique chronologique qui domine le corpus et nous permet d'interpréter l'arbre. De de Gaulle (1958-1969) à Mitterrand 2 (1981-1995) le discours politique français perd de ses substantifs, adjectifs et déterminants au profit des verbes, des pronoms et des adverbes. En conséquence, Giscard d'Estaing, chronologiquement situé entre de Gaulle et Mitterrand, tient un discours équilibré, partagé entre le nominal et le verbal ²⁴.

En revanche, le discours de Chirac ne laisse pas d'intriguer. Aucune explication ne semble pouvoir rendre compte de sa situation centrale sur l'arbre. L'ordre chronologique n'est pas respecté puisque Chirac, sur l'arbre, est tout aussi proche (ou éloigné) du début de la ^ve République que de la période récente. L'ordre politique semble également contredit puisque Chirac se trouve proche également (ou éloigné) de Mitterrand que de de Gaulle. Jacques Chirac a un discours « moyen », tempéré, du juste milieu ; pour cette raison, il est bien difficile à analyser.

Reste pour finir l'impression historique d'ensemble : le discours présidentiel, de de Gaulle à Mitterrand ou Chirac, de 1958 à 1995 ou 2002, illustre à nos yeux la mort du politique. Les discours des présidents, en perdant leurs substantifs, perdent de leur substance politique. Par la répétition des verbes (le plus souvent énonciatifs) et des pronoms (dominés par

22. Pour une interprétation aboutie : MAYAFFRE, D., ouvrage à paraître 2004.

23. BRUNET, É., 1981.

24. Dans le détail, Giscard utilise énormément de noms (comme de Gaulle ou Pompidou), mais il ne répond pas à toutes les exigences d'un discours nominal en utilisant assez peu d'adjectifs ou de déterminants. De même, il emploie beaucoup de verbes, mais refuse les adverbes et les pronoms.

le « je » du locuteur), ils ne sont plus que la mise en scène d'un homme et d'une promesse d'action. Les réflexions sur la finalité du Politique laissent leur place aux modalités de l'agir politique (la redondance des adverbes chez J. Chirac est à ce titre très significative). Le discours politique français devient au fil du temps pratique ou pragmatique, mais en bannissant les noms et les concepts, il renonce non seulement à l'idéologie et aux idéaux, mais, plus gravement, aux idées ²⁵.

2. Arbre et pomme de discorde : la cohabitation (1997-2002), et les racines de la défaite de L. Jospin

L'analyse arborée permet de traiter des corpus lourds et complexes. Le corpus de la cohabitation Jospin/Chirac (1997-2002) que l'on se propose d'analyser maintenant est composé de tous les discours que Chirac et Jospin ont prononcés ou écrits en France entre 1997 et 2002, soit quelque 2 100 discours et 2 400 000 mots. Ce corpus est segmenté politiquement et chronologiquement en douze sous-parties (et non plus seulement en six comme précédemment) dont on voudrait voir les rapports (Tableau 3).

Tableau 3. *Le corpus de la cohabitation Chirac/Jospin (1997-2002)*

<i>Textes</i>	<i>Nombre de discours</i>	<i>Nombres d'occurrences</i>	<i>Nombre de vocables différents</i>
Chirac1997	73	104 553	8 902
Chirac1998	92	183 820	11 910
Chirac1999	95	176 352	10 989
Chirac2000	135	251 570	12 917
Chirac2001	107	219 236	12 110
Chirac2002	43	209 694	11 638
<i>Total partiel (1)</i>	<i>595</i>	<i>1 145 225</i>	<i>24 884</i>
Jospin1997	39	56 229	5 935
Jospin1998	79	190 455	11 736
Jospin1999	135	225 158	12 915
Jospin2000	153	300 308	15 035
Jospin2001	150	276 747	14 077
Jospin2002	74	227 700	12 381
<i>Total partiel (2)</i>	<i>630</i>	<i>1 276 597</i>	<i>26 593</i>
Total	1 225	2 421 822	38 849

25. MAYAFFRE, D., 2004.

Face à ce corps de discours, ainsi segmenté, deux types de questionnements peuvent être formulés. Premièrement, y a-t-il une identité linguistique chiraquienne et une identité linguistique jospinienne ? Les deux hommes tiennent-ils le même langage ou bien des signes linguistiques forts et originaux marquent-ils leurs discours respectifs ? Au fond, y-a t-il une distance entre les discours du Président et ceux du Premier ministre ? Deuxièmement, quels sont l'ampleur et le rythme de l'évolution chronologique dans le corpus ? Les discours sont-ils identiques à eux-mêmes entre 1997 et 2002 ou changent-ils au fil des ans de manière continue et/ou chaotique ? Finalement, quelle est la distance entre un discours de 1997, un de 2000 et un de 2002 ?

Surtout, les deux types de questionnement peuvent être croisés. Quel est le facteur qui apparaît le plus prégnant dans le corpus ? Le facteur politique ou le facteur chronologique ? L'identité politico-linguistique de Chirac et Jospin résiste-t-elle aux contraintes imposées par le temps ? Les discours de Chirac — quelle que soit la conjoncture — s'opposent-ils toujours à ceux de Jospin ? Ou, au contraire, le temps, avec les thèmes qu'il charrie et impose, commande-t-il les discours ? Dans cette hypothèse, les discours d'une même année se ressembleraient toujours par-delà l'identité politique différente de leur locuteur.

Par la même procédure que celle décrite précédemment, un indice de distance intertextuelle, calculé selon l'ensemble du vocabulaire employé dans les discours, a été établi pour aboutir au tableau matriciel suivant (Tableau 4).

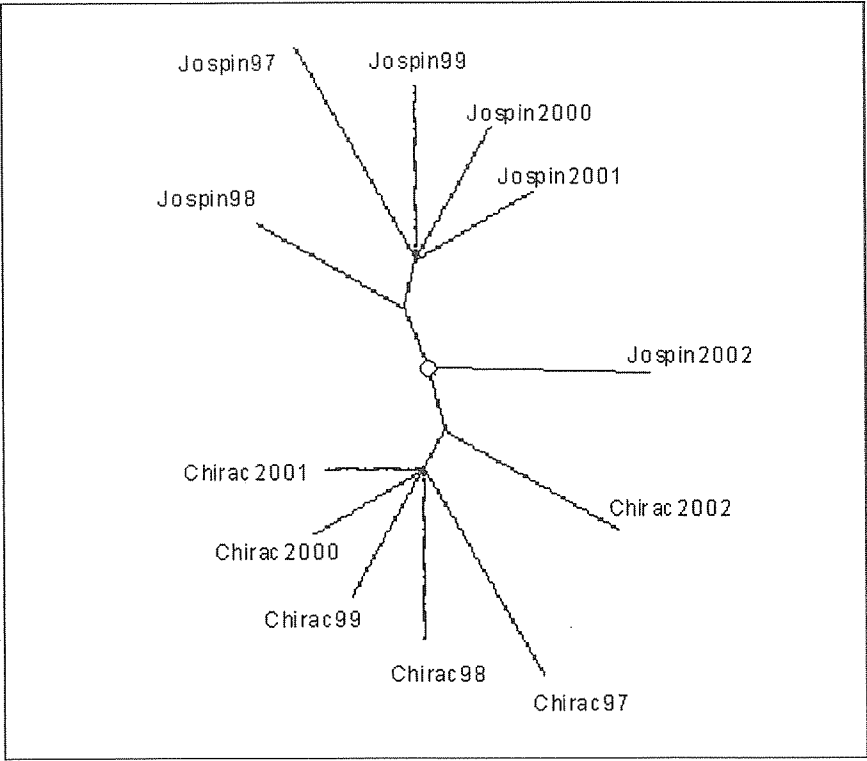
L'analyse arborée prouve alors toute sa puissance en donnant une représentation fine et précise de cette grande matrice (Figure 5).

D'un tel arbre, trois conclusions socio-linguistiques peuvent être tirées. D'abord, il existe une individuation linguistique réelle entre Chirac et Jospin. L'algorithme dessine nettement deux branches opposées. Sur l'une se distinguent toutes les ramifications de Chirac, sur l'autre toutes les ramifications de Jospin. (On enregistre une seule exception, nous y reviendrons, le discours de Jospin 2002 qui forme à lui tout seul une branche indépendante.) Le temps ne perturbe que très peu cette dichotomie politique. Après beaucoup d'autres expériences, avec des outils variés, sur des corpus politiques divers, nous pouvons à nouveau affirmer que le facteur individuel et politique semble toujours prépondérant dans le discours politique. Lorsque l'on croise plusieurs facteurs — l'identité

Tableau 4. Indices des distances entre textes selon le vocabulaire dans le corpus Cohabitation (1997-2002)

	C.97	C.98	C.99	C.00	C.01	C.02	J.97	J.98	J.99	J.00	J.01	J.02
C.97	0	0,110	0,116	0,113	0,124	0,145	0,197	0,167	0,183	0,179	0,173	0,168
C.98	0,110	0	0,089	0,094	0,098	0,119	0,182	0,148	0,168	0,167	0,162	0,151
C.99	0,116	0,089	0	0,088	0,092	0,123	0,180	0,151	0,160	0,163	0,160	0,155
C.00	0,113	0,094	0,088	0	0,083	0,118	0,171	0,142	0,155	0,150	0,147	0,146
C.01	0,124	0,098	0,092	0,083	0	0,107	0,166	0,143	0,150	0,150	0,139	0,148
C.02	0,145	0,119	0,123	0,118	0,107	0	0,178	0,150	0,168	0,168	0,161	0,139
J.97	0,197	0,182	0,180	0,171	0,166	0,178	0	0,133	0,123	0,124	0,126	0,175
J.98	0,167	0,148	0,151	0,142	0,143	0,150	0,133	0	0,114	0,118	0,113	0,131
J.99	0,183	0,168	0,160	0,155	0,150	0,168	0,123	0,114	0	0,094	0,096	0,158
J.00	0,179	0,167	0,163	0,150	0,150	0,168	0,124	0,118	0,094	0	0,092	0,152
J.01	0,173	0,162	0,160	0,147	0,139	0,161	0,126	0,113	0,096	0,092	0	0,144
J.02	0,168	0,151	0,155	0,146	0,148	0,139	0,175	0,131	0,158	0,152	0,144	0

Figure 5. Analyse arborée selon le vocabulaire du corpus Cohabitation (1997-2002)



politique du locuteur, la chronologie, le genre des discours, etc. — c'est toujours l'identité du locuteur — son appartenance politique et son statut — qui apparaît primordiale²⁶.

Ensuite, sur chacune des branches, c'est-à-dire localement sur l'arbre, le graphe se montre très précis, notamment dans la description de la chronologie et de l'évolution du discours durant la période.

Il est très intéressant de remarquer comment le discours de Chirac 2002, par une sous-branche particulière, change en partie de nature et cherche à gommer son identité chiraquienne originelle. Il ne fait aucun doute que la campagne électorale se fait sur des thèmes jusqu'alors inexploités, avec d'autres mots que ceux des discours antérieurs. Pour des raisons électorales, Chirac semble inventer, en 2002, un autre discours. Néanmoins — et la précision prendra tout son sens plus loin —, ce discours de campagne, en grande partie original, reste clairement affilié au parler de Jacques Chirac : il reste, sur l'arbre, sur la même branche maîtresse chiraquienne avec les autres discours du président.

Du reste, chez Chirac, cette évolution très marquée et un peu exceptionnelle du discours de 2002 ne doit pas nous masquer l'évolution chronologique progressive et cohérente de l'ensemble du discours. Sur l'arbre, dans le bouquet des textes de Chirac entre 1997 et 2001, on remarque que la distance (c'est-à-dire l'originalité) tend à se réduire au fil du temps : les ramifications sont en effet de plus en plus courtes, au fur et à mesure que s'égrainent les années. En d'autres termes — et l'observation attentive des indices du tableau peut le confirmer — le discours de Chirac en 1997 est le plus éloigné du discours moyen (c'est-à-dire en grande partie, par effet de contraste dans notre corpus, du discours de Jospin), puis en 1998 le président perd en partie de son originalité comme l'atteste une ramification plus courte en 1999, il en perd encore davantage, etc.

Du côté Jospin, la même tendance est observée. À mesure que le temps s'écoule, Jospin se rapproche d'un discours moyen (c'est-à-dire dans notre corpus du discours chiraquien). Une seule exception perturbe légèrement ce bel ordonnancement chronologique : le discours de 1998 est très précocement édulcoré se démarquant de ses frères pour se rapprocher

26. MAYAFFRE, D., 2002, pp. 517-529.

27. Une explication de la particularité de l'année 1998 durant la gouvernance de Lionel Jospin sera donnée dans D. MAYAFFRE, 2004.

du centre topologique de l'arbre ²⁷. Mais, au-delà de cette exception, c'est le discours de 1997 qui est le plus typé ou le plus marginal (ramification la plus longue), puis celui de 1999, puis celui de 2000, etc.

Ainsi la conclusion socio-linguistique est évidente : les deux hommes déteignent l'un sur l'autre, et leurs discours tendent à se rapprocher. L'effet émollient de la cohabitation où le clivage gauche/droite est estompé est ainsi démontré. Sans jamais tenir le même discours, Jospin et Chirac semblent perdre petit à petit, entre 1997 et 2002, leur identité propre pour employer de plus en plus un vocabulaire commun.

Enfin, le sel de l'arbre de la cohabitation se trouve dans la situation particulière du texte Jospin2002. Que les discours des deux hommes tendent à se rapprocher a déjà été dit. Que, symétriquement à celui de Jospin, le discours de campagne de Chirac soit particulier et opportuniste a déjà été souligné. Mais la localisation de Jospin2002 dépasse tous ces constats pour expliquer la défaite du Premier ministre aux élections présidentielles.

Il ne s'agit plus d'une évolution du discours mais d'une rupture linguistique étonnante. Dans le clivage linguistico-politique Jospin/Chirac dont elle donne une belle représentation, l'analyse arborée n'arrive plus à reconnaître et à classer Jospin2002. Le discours de campagne du Premier ministre se trouve aussi éloigné du discours de Chirac que des siens propres durant les années précédentes. En 2002, Jospin tient un discours à part, original, mutant, qui ne s'apparente pas à celui qu'il a tenu précédemment ²⁸. La personnalité linguistique (sinon politique) de l'émetteur, que les Français avaient appris à reconnaître durant cinq ans, est totalement brouillée. Pour cette raison, le discours devient inaudible, particulièrement pour l'électorat de gauche. Est-ce la faute des conseillers en communication ou des plumes du candidat ? Est-ce la faute des médias et de ses adversaires qui ont su entraîner Jospin sur des thématiques qui lui étaient étrangères ? En tout cas, les mots que le Premier ministre candidat se met alors à utiliser ne lui appartiennent pas. En 2002, Jospin vocalise un discours étranger, « inidentifiable ». Dans ces conditions, il avait peu de chance de se faire entendre.

28. On remarquera même de manière symptomatique que le texte le plus éloigné (0.175) du texte Jospin 2002 n'est pas un texte de Chirac mais le texte Jospin 1997 : dans le corpus, les discours qui ressemblent le moins aux discours de Jospin 2002 sont les discours de Jospin tenus 5 ans auparavant.

*

L'analyse arborée permet une représentation globale mais détaillée, synthétique mais fine, d'une matrice de données. Dans cet article, ces données ont toujours été des indices de distances ²⁹ : distances entre les textes des six septennats de la v^e République (1958-2002), distances entre les douze textes de la cohabitation Jospin/Chirac (1997-2002). Aussi notre démonstration a-t-elle paru facilitée par le jeu de mots autour du terme « distance » parfois entendu dans le sens statistique (= calcul) parfois entendu dans le sens géographique (= mesure). Ce jeu de mots n'est pas fantaisie. Il est dans l'esprit même de l'analyse arborée.

La distance intertextuelle entre les textes (t1, t2, t3,...) d'un corpus, exprimée après calcul par un indice mathématique, est représentée le plus exactement possible par une distance géographique entre les feuilles-textes (f1, f2, f3, etc.) de l'arbre. Sans donc connaître tous les tenants et les aboutissants de la topologie mathématique, la lecture d'un graphe arboré et son interprétation peuvent difficilement être plus intuitives. Aussi naturellement que l'on constate sur une carte routière que Marseille est plus éloignée de Paris que de Lyon, il a été possible de lire sur le premier arbre produit dans cette étude que les discours de de Gaulle étaient plus éloignés de ceux de Mitterrand que de ceux de Giscard. Rappelons simplement qu'en sus de la distance géographique entre les textes, l'analyse arborée présente par la structure de l'arbre une dimension classificatoire précieuse en regroupant les textes apparentés et ceux qui ne le sont pas. En filant la comparaison, on peut dire qu'au-delà de la distance à vol d'oiseau entre deux villes, le fait qu'elles soient ou non sur un même axe (branche) autoroutier ou desservies par le même nœud ferroviaire prend tout son sens pour le voyageur.

Depuis A. Prost, en histoire, lorsqu'il s'agit de croiser plusieurs variables, de traiter un tableau rectangulaire et de situer par exemple des textes selon le vocabulaire utilisé (n textes x z mots), c'est l'analyse factorielle des correspondances (AFC) qui est pratiquée ³⁰. L'ambition de l'analyse arborée est au fond identique à l'AFC même si le mécanisme calculatoire n'est pas similaire et la représentation graphique différente. Du reste, toutes les expériences montrent que les deux méthodes donnent des résultats convergents et complémentaires ³¹.

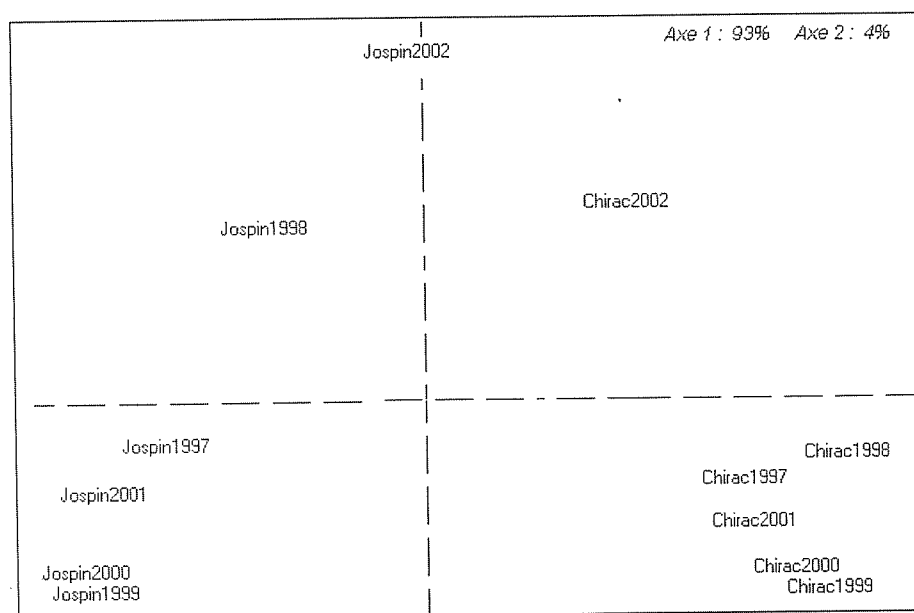
29. Mais d'autres données peuvent être traitées comme un simple tableau de fréquences.

30. PROST, A., 1974 et d'autres articles du même auteur.

Dans ce cadre, précisons que le principal atout de l'analyse arborée est de représenter dans un seul mouvement, c'est-à-dire sur les deux seules dimensions du plan où se dessine l'arbre, toutes les données de la matrice, lorsque l'AFC demande la consultation successive du premier, deuxième, troisième, quatrième, etc. axes factoriels pour rendre compte de toutes les nuances. Le deuxième avantage vient des vertus de la classification locale. Si les AFC indiquent, avec le succès que l'on sait, les grandes tendances et oppositions, toujours très nettes sur l'axe 1 et l'axe 2, localement dans le nuage de points, elles n'ordonnent que mal, entre eux, les éléments assez proches, à moins de recourir par des lectures successives difficiles pour le non spécialiste aux axes marginaux. L'arborée, en revanche, définit de grandes classes de textes apparentés et les oppose entre elles. Puis, à l'intérieur de ces classes, localement sur les branches, elle distribue les textes, éventuellement selon des regroupements secondaires, toujours selon une distance significative.

Ainsi l'arbre de la cohabitation Chirac/Jospin que nous avons produit (Figure 5) peut être mis avantageusement en parallèle avec une AFC issue du traitement des mêmes données par le logiciel *Hyperbase* (Figure 6).

Figure 6. AFC du corpus *Cohabitation* (1997-2002) selon le vocabulaire



31. Par exemple, D. MAYAFFRE, 2002.

Le schème général de l'AFC est identique à celui de l'arbre. La méthode s'en trouve validée et les conclusions linguistico-politiques, en trois points, sont confirmées : il existe bien une identité lexicale chiracienne *versus* jospinienne apparente sur l'axe 1 (axe horizontal) ; les discours de Chirac2002 ou Jospin1998 apparaissent comme particuliers dans leur famille respective ; enfin, Jospin2002 n'est plus seulement particulier mais totalement atypique, exactement central sur l'axe 1 et excentré sur l'axe 2 (axe vertical).

Localement, en revanche, sur l'AFC, les concentrés de points (Chirac1997, Chirac1998, Chirac1999, Chirac2000, Chirac2001 et Jospin1997, Jospin1999, Jospin2000, Jospin2001) sont peu précis et difficiles à interpréter sans la consultation de l'axe 3 voire de l'axe 4. L'analyse arborée, elle, nous avait pourtant permis de constater, du premier coup d'œil, l'évolution qui avait cours durant la cohabitation, en soulignant — par la dimension de plus en plus courte au fil des années des ramifications menant aux textes — comment les discours du Président et du Premier ministre perdaient, dans la durée, de leur originalité respective pour tendre vers leur confusion.

Ainsi après des exemples probants sur des corpus littéraires ³², l'analyse arborée a montré son efficacité sur deux lourds corpus politiques en nous aidant à cerner le discours présidentiel sous la v^e République et particulièrement le caractère « moyen », pas plus gaullien que mitterrandien, du discours de J. Chirac, à comprendre le mécanisme de la cohabitation (uniformisation, à terme, des discours ; déteinte du parler des cohabitants) et à mesurer le drame de L. Jospin qui au bout d'un processus politique et linguistique de cinq années entre en campagne, en 2002, avec un discours mutant, littéralement inclassable, étranger à lui-même et inaudible pour son électorat.

32. Par exemple, M. KASTBERG SJÖBLOM, 2002.

Bibliographie

- BARTHÉLEMY, Jean-Pierre & LUONG, Xuan, « Représenter les données textuelles par les arbres », *JADT* 98, Nice, 1998, pp. 49-73.
- BRUNET, Étienne, *Le vocabulaire français de 1789 à nos jours*, Paris-Genève, Champion-Slatkine, 1981.
- « Une mesure de la distance intertextuelle : la connexion lexicale », *Revue informatique et statistique dans les sciences humaines, Le nombre et le texte*, hommage à Étienne Evrard, Liège, 1988, n° 1 à 4, pp. 81-116.
- « Peut-on mesurer la distance entre deux textes ? », *Corpus*, n° 2, 2003, pp. 47-70.
- Corpus*, « Distances entre textes », n° 2, 2003, à paraître.
- JACQUART, Pierre, « Nouvelles recherches sur la distribution florale », *Bull. Soc. Vand. Sci. Nat.*, n° 44, 1908.
- JUILLARD, Michel, « Entre le silence et le bruit : les arbres », *Travaux du cercle linguistique de Nice*, n° 16, 1994, pp. 3-24.
- KASTBERG SJÖBLOM, Margareta, *L'écriture de J.M.G. Le Clézio, une approche lexicométrique*, Thèse de doctorat, Nice, novembre 2002.
- LABBÉ, Cyril et Dominique, « La distance intertextuelle », *Corpus*, n° 2, 2003, pp. 95-118.
- LABBÉ, Dominique & HUBBERT, Pierre, « La connexion des vocabulaires », *JADT* 1998, Nice, 1998, pp. 361-370.
- LABBÉ, Dominique & MONIÈRE, Denis, *Le discours gouvernemental. Canada, Québec, France (1944-2000)*. Paris, Champion, col. Lettres numériques, 2003, à paraître.
- LUONG, Xuan, *Méthodes d'analyse arborée. Algorithmes, applications*, Thèse d'État, 1988, Paris v.
- LUONG, Xuan, « L'analyse arborée des données textuelles : mode d'emploi », *Travaux du cercle linguistique de Nice*, n° 16, 1994, pp. 25-43.
- LUONG, Xuan & MELLET, Sylvie, « Les calculs multidimensionnels au service de l'analyse syntaxique diachronique », *JADT* 1995, Rome, CISU, 1995, vol. II, pp. 281-289.
- MAYAFFRE, Damon, « Discours politique, genres et individualisation socio-linguistique », *JADT* 2002, INRIA-IRISA, Saint-Malo, 2002, pp. 517-529.
- *Jacques Chirac, le président de la parole. Étude du discours présidentiel français (1958-2003)*, Paris, Champion, 2004, (à paraître).
- MULLER, Charles, *Principes et méthodes de statistique lexicale*, Paris, Hachette, 1977, rééd Champion, 1992.
- PINCEMIN, Bénédicte, « Similarités texte-textes. Expérience d'une application de diffusion ciblée et propositions », *Matemáticas y Tratamiento de Corpus*, (s.n.), Logroño, Fundacion San Millán de la Cogolla, 2002, pp. 35-52.
- PROST, Antoine, *Vocabulaire des proclamations électorales de 1881, 1885 et 1889*, Paris, PUF, 1974.